Cardiovascular Disease Predictor

Daksh Jain[1], Vardan Yadav[2], Manavi Kumari[3], Kundan Chandravansi[4], Gurakonda Konda Reddy[5], Jeba Nega Cheltha[6]
{jaindaksh58@gmail.com[1], vardanyadav36@gmail.com[2], manavi.16kumari@gmail.com[3], kundanrkt@gmail.com[4], gkondareddy18@gmail.com[5], jeba.25111@lpu.co.in[6]}

School of Computer Science and Engineering, Lovely Professional University, Punjab, India[1,2,3,4,5,6]

Abstract. Heart disease is a very common disease that kills many people every year regardless of age. So, machine learning is used to predict if a person has cardiovascular disease. The classification models used here are KNN, Ridge Classifier, Decision Tree, Random Forest and GaussianNB. The dataset used is the Cleveland heart dataset. In this paper, we firstly pre-process the data. After that, the dataset is branched into two parts, the training and the testing parts. After that, we fit the data in the models and get an initial idea prediction. Then, we apply hyperparameter tuning to increase the accuracy of the models. After hyperparameter tuning, we find that KNN performed best with the roc_auc score of 93.2%. This accuracy is an improvement over previous work. A web application using Flask is also created to provide GUI to users.

Keywords: Cardiovascular disease prediction, GaussianNB, KNN, Random Forest, DecisionTree, RidgeClassifier.

1. Introduction

In this era of Covid, we are witnessing a scarcity of doctors. A lot of people are suffering without any help being provided to them and lack of any proper medical equipment to measure such disease. Heart disease can be explained as something which affects your heart in a number of ways. Heart disease is a broader term under which many disease are present such as Coronary artery disease, arrhythmia, heart valve disease and heart failure. Cardiovascular disease refers to a condition involving narrowing or blocking of blood vessels in our body steering to various condition such as chest pain stroke and heart attack. Our paper pit multiple machine learning classification models against each other so as to get an idea which model is best suited for the task and uses hyperparameter tuning to further increase the final accuracy. This paper proposes an online web application providing a fast and reliable prediction on heart condition. It then processes those provided details to check whether the person has the cardiovascular disease or not. The system can be used to get an initial idea on the condition of the patient. For this task, we have used the dataset which is the Cleveland heart disease dataset from UCI repository[1] to understand which model would be the best machine learning model for our task and then apply that model to predict the outcome and show it to the user.

2. Related Work

A. Rajdhan et. al. [2], has written a paper which helps us in predicting the status of person's heart condition. In this paper they have used various data mining techniques like Random Forest, Logistic Regression, Decision Tree and Naive Bayes. After Comparing the results from all the different models, they came to the conclusion that Random Forest perform best with the accuracy score of 90.16%.

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318172

P. Anbuselvan [3] in her paper, she compared various supervised learning models such as Random Forest, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Logistic Regression, Naïve Bayes, and also the ensembling technique of XGBoost to find out the best algorithm for the purpose. She found Random Forest to be the most accurate model with 86.89% accuracy.

H. Meshref [4] trained different prominent machine learning techniques: Random Forests, Support Vector Machines, Naïve Bayes, Decision Trees and also a deep learning technique Artificial Neural Networks. In his study he found Artificial Neural Network to be the model with the best accuracy of 84.25%.

R. Magar et. al. [5], designed a system that helps in discovering the hidden patterns in the dataset by using various data mining techniques, and forecasting if a person is diagnosed with cardiovascular disease. In their paper they used models namely Naïve Bayes, Logistic Regression, Support Vector Machine and Decision Tree. They later discovered that logistic Regression provided them with the best result with accuracy of 82.89%.

P. K. Bhunia et. al. [6], constructed a Heart Disease Prediction System (HDPS) comparing a horde of Machine Learning models such as K Nearest Neighbor, Support Vector Machine, Random Forest Classifier, Logistic Regression and Decision Tree models to project the risk index of heart. Finally, the results they obtained provided them with the information that the Support Vector Machine and Random Forest Classifier both got an accuracy of 90.32% which is the highest among all other models they used.

A.N. Repaka et. al. [7], performed different data mining techniques to achieve a Smart heart disease prediction using techniques namely Sequential Minimal Optimization (SMO), Bayesian Net (BN), Multilayer Perception (MLP) and Naïve Bayesian. After performing these techniques, they got the result which showed Naïve Bayesian performed best with an accuracy of 89.77%.

R. Goel [8] used feature selection technique to find out that attributes such as cholesterol (chol), sex, blood pressure(bp), chest pain (cp) to be the most useful. Moving forward with these attributes she used a tally of 6 algorithms of Machine Learning including KNN, Logistic Regression, Decision Tree, Naïve Bayes ,Random Forest and SVM to get the prediction on whether the person have cardiovascular disease or not. Using Confusion Matrix as a metric she came to the conclusion that SVM helmed the crown with an accuracy of 86%.

H. Jindal et. al. [9], performed heart disease analysis using distinct machine learning model like Logistic regression, KNN, and Random Forest Classifier. They concluded that KNN perform best for their dataset with a calculated accuracy of 88.52%

A.U. Haq et. al. [10], used 3 different feature selection techniques to get the best attributes, 7 plethora machine learning algorithms such as Logistic regression, KNN, Decision Tree, ANN, Naïve Bayes, SVM and random forest.7 different classifier performance evaluation metrics like Matthews' correlation coefficient, execution time, classification accuracy, specificity, sensitivity and finally cross validation method to validate their prediction. After performing all the task, they came to the conclusion that logistic regression achieves the best accuracy of 89%.

S. Mohan et. al. [11], used feature selection and an enhanced version of the machine learning model with different combination of features and classification technique which is the hybrid of linear model and random forest classifier to get the prediction of heart disease and they got the accuracy of 88.7%.

N. Basha et al. [12], applied different machine learning models on a Kaggle dataset of heart disease using the models Decision Tree, SVM, KNN, Random Forest and Naïve Bayes. They also performed exploratory data analysis. After using all the models, they got the best accuracy from KNN which is 85%.

R. Shrestha et al. [13], researched a lot of diverse ML models such as Artificial Neural Network, Decision Tree, KNN, Naïve Bayes and Artificial Neural Network. She came to the conclusion that naïve Bayes perform the best with accuracy of 88.163%.

3. Methodology

In this paper we analyse multiple machine learning models, the algorithms used here are GuassianNB, Ridge Classifier, K nearest neighbors (KNN) [14], Decision Tree [15] and Random Forest [16] which will help us determine the algorithm with the highest accuracy to get the prediction on the person having Cardiovascular disease or not. In this paper we also give an insight of the different research papers published on this topic. The data which we will work on is Cleveland heart dataset. To get the best model for this task firstly we need to clean and organize the data properly which is done using the data pre-processing. Secondly, we draw graph of the data to see any pattern the data follows. After getting the idea about the data, we split the data in training dataset, on which we perform the analysis, and the testing dataset, on which we test the analysis we did on the training set. Then we apply the different machine learning models which are Decision Tree, GuassianNB, Ridge Classifier, K nearest neighbors (KNN) and Random Forest to get the initial result. After that we perform the Hyperparameter Tuning [17] to maximize the accuracy of each model and finally, we get the best accuracy for a system. The performance metric used here is the roc_aur score.

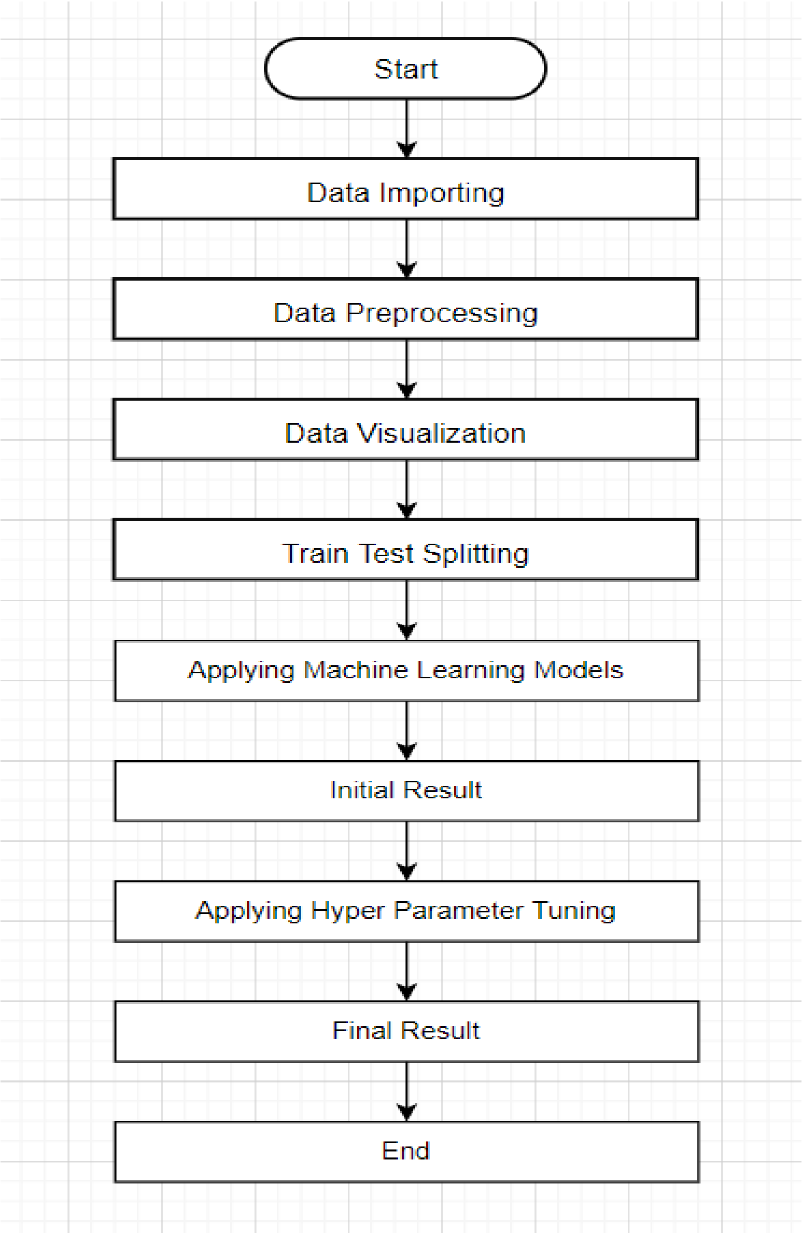

Fig. 1. Cardiovascular Disease Prediction Flowchart

4. Algorithms & Performance Metric

4.1. Machine learning Algorithms

4.1.1 KNeighbors classifier. A supervised learning classification technique used in analysing the distance of various points of data to all the different groups of data present and categorizing which data point will belong to which data group based on the closeness of the data points to groups. In addition to solving the classification problems, it also helps us in solving the regression problems. It is also called a lazy learner because of the fact that it does not learn but rather uses the previous result to give out the prediction

4.1.2 Decision Tree Classifier. A method used for both regression and classification and is a form of supervised learning, this classifier basically sets a number of rules and then makes decision based on them. The intuition behind this type of model is that it branches into two parts of yes and no and continues this process till the isolation of every data point occurs. This process creates a tree like structure with each yes, no question in the form of a node of that tree. We start from the top and chooses for each leaf node until we reach our final result**.**

4.1.3 Random Forest Classifier. It is a part of supervised learning technique. This can be apply for Regression problems as well as Classification Problems in ML. This contain many decision trees on several sub-category of given data set and take mean to get best accuracy. The greater the number of trees in the forest the more and better will be the accuracy and it also prevents the problem of overfitting of the data set.

4.1.4 Ridge Classifier. A classification method based on Ridge Regression method which essentially starts with converting the labelled data points into -1 and +1.The target value is the prediction with highest value and multi-output regression is used in the case of multi-class data. It helps in both classification and regression. It can be further defined as a model whose existence is very similar to the linear regression model but a subtle difference is that it has a very low variance compared to the normal Linear Regression algorithm and a much higher bias. This is because this model is generally used for long term applications where the changes in bias and variance really shines. Finally the accuracy of this model is calculated using the confusion matrix.

4.1.5 GaussianNB. Based on the Naive Bayes algorithm, this classifier is based on the assumption that the data in each label is obtained using a Gaussian (normal) distribution of the data. We need two things, i.e., the standard deviation and the mean and distribution of the data values in each individual label, to interpret them and finally fit the model. The preferred data here is continuous data.

4.2 Performance Metric

The metric we used to get the accuracy for the different models is the ROC_AUC score . The Receiver Operator Characteristic (ROC) curve can be explained as the probability curve which maps the false positive (FPR) against True positive (TPR) at numerous threshold values and distinguishes the ‘signal’ from the ‘noise’. The Area under the Curve (AUC) is fundamentally used to find a distinction between classes and also provide a concise report of ROC curve. Greater the performance of the algorithm at determining the difference between the positive and negative classes, higher is the accuracy.

5. Result and Discussion

The different models we used were RidgeClassifier which gave an initial auc_roc score of 84.5%, KNeighbors Classifier which gave a score of 87.3%, GaussianNB Classifier which gave a score of 81%, DecisionTree Classifier which gave a score of 81%, RandomForest Classifier which gave a score of 83.5%.

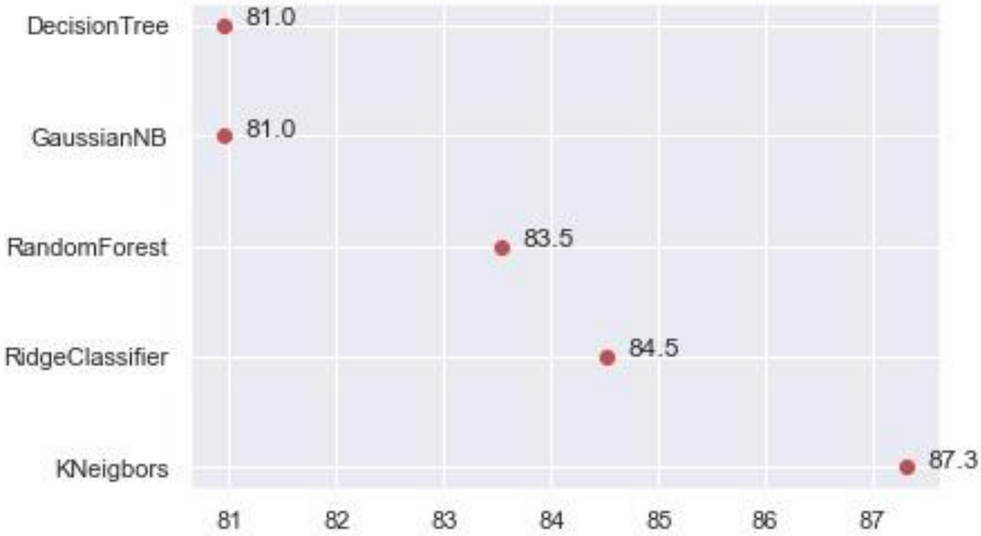

Fig. 2. Model comparison before Hyperparameter Tuning

After we performed the Hyperparameter tuning, our models saw a good increase in the auc_roc score which is, RidgeClassifier which gave the increased score of 90.8%, KNeighbors Classifier which gave the increased score of 93.2%, GaussianNB Classifier which gave the increased score of 90.6%, DecisionTree Classifier which gave the increased score of 84.5%, and finally RandomForest Classifier which gave the increased score of 92.7%. After performing hyperparameter tuning, the accuracy for all the models increased drastically but the highest accuracy we got is from the KNeighbors Classifier, so we decided to use it for the prediction.

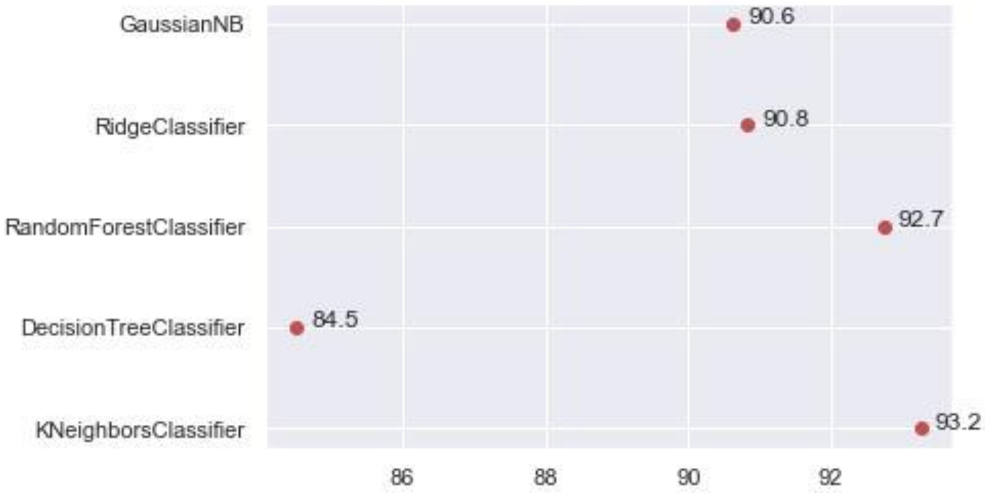

Fig. 3. Model Comparison after Hyperparameter Tuning

6. Conclusion & Future Work

The rising mortality rate of patients with heart disease has increased the need for a sudden solution. Keeping this problem in mind, a system is urgently needed to help reduce this mortality rate. The motivation for the study was to find the most efficient Machine Learning model for the correct prediction of whether the person is suffering from the cardiovascular disease or not and to improve the prediction rate from the already existing models and researches.

A total of five machine learning models namely KNeighbor Classifier, RidgeClassifier, DecisionTree Classifier, RandomForest Classifier, GaussionNB Classifier. The dataset we used is Cleveland Data set which has 14 attributes. Initially we got the best accuracy from KNeighbor Classifier which was about 87.3%. After performing hyperparameter tuning, we still got the best accuracy from the KNeighbor Classifier but it jumped to 93.2% from the initial 87.3% with the efficient values after tuning the parameters, in hyperparameter tuning, coming out to be n_neighbors = 21 and metric = ‘manhatten’. This process is both time efficient and cost efficient and there are many more Machine Learning models that can be used for further increasing the accuracy for this dataset in the future. This accuracy we achieved can further be improved with the help of feature selection [18] and analyzing the data further which will help us in saving more lives.

References

[1] Magar, R., Memane, Raut, S., Rupnar, V. S.: Heart Disease Prediction using Machine Learning. Journal of Emerging Technologies and Innovative Research (JITER), Vol. 7, Issue 6 (2020).

[2] Rajdhan, A., Sai, M., Agarwal, A., Ravi, D., Ghuli, P.: Heart Disease Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 04 (2020).

[3] Anbuselvan, P.: Heart Disease Prediction using Machine Learning Techniques. International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 11 (2020).

[4] Meshref, H.: Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12 (2019).

[5] Magar, R., Memane, Raut, S., Rupnar, V. S.: Heart Disease Prediction using Machine Learning. Journal of Emerging Technologies and Innovative Research (JITER), Vol. 7, Issue 6 (2020).

[6] Bhunia, P. K., Debnath, A., Mondal, P., Monalisa, D. E., Ganguly, K., Rakshit, P.: Heart Disease Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 11 (2021).

[7] Repaka, A. N., Ravikanti, S. D., Franklin, R. G.: Design and Implementing Heart Disease Prediction Using Naives Bayesian. Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 292-297 (2019).

[8] Goel, R.: Heart Disease Prediction Using Various Algorithms of Machine Learning. Proceedings of the International Conference on Innovative Computing & Communication (ICICC) (2021).

[9] Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P.: Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering (2021).

[10] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., Sun, R.: A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mobile Information Systems, vol. 2018, Article ID 3860146 (2018).

[11] Mohan, S. K., Thirumalai, C., Srivastava, G.: Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, vol. 7, pp. 81542-81554 (2019).

[12] Basha, N., Ashok Kumar, P. S., Krishna, C. G., Venkatesh, P.: Early Detection of Heart Syndrome Using Machine Learning Technique". 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 387-391 (2019).

[13] Shrestha, R., Chatterjee, J. M.: Heart Disease Prediction System Using Machine Learning. LBEF Research Journal of Science, Technology and Management, Vol. 1, Issue 2 (2019).

[14] Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P.: Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering (2021).

[15] Rajdhan, A., Sai, M., Agarwal, A., Ravi, D., Ghuli, P.: Heart Disease Prediction using Machine Learning. International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 04 (2020).

[16] Jindal, H., Agrawal, S., Khera, R., Jain, R., Nagrath, P.: Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering (2021).

[17] Budholiya, K., Shrivastava, S. K., Sharma, V.: An optimized XGBoost based diagnostic system for effective prediction of heart disease. Journal of King Saud University (2020).

[18] Khemphila, A., Boonjing, V.: Heart disease Classification using Neural Network and Feature Selection. IEEE Xplore, pp. 406-409 (2011).